

BEST AVAILABLE COPY

日 本 国 特 許 庁

JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 7 月 1 7 日
Date of Application:

出 願 番 号 特 願 2 0 0 3 - 2 7 5 9 8 3
Application Number:

ST. 10/C] : [J P 2 0 0 3 - 2 7 5 9 8 3]

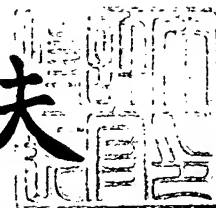
願 人
Applicant(s): 日本電気株式会社
 大瀧 慈
 社団法人バイオ産業情報化コンソーシアム

CERTIFIED COPY OF
PRIORITY DOCUMENT

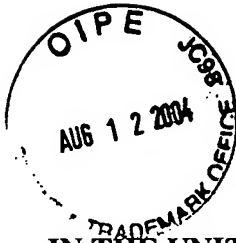
2 0 0 3 年 1 0 月 2 9 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 0 8 9 5 5



PATENT APPLICATION
IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of

Docket No: Q79665

Masataka ANDOH, et al.

Appln. No.: 10/766,011

Group Art Unit: 1631

Confirmation No.: 2008

Examiner: Unknown

Filed: January 29, 2004

For: **SYSTEM, METHOD, AND PROGRAM FOR ESTIMATING GENE EXPRESSION
STATE, AND RECORDING MEDIUM THEREFOR**

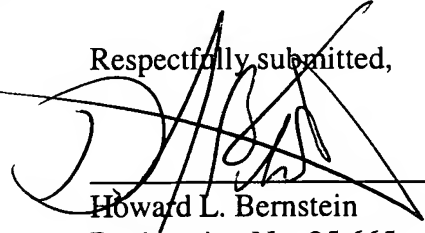
SUBMISSION OF PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith is a certified copy of the priority document on which a claim to
priority was made under 35 U.S.C. § 119. The Examiner is respectfully requested to
acknowledge receipt of said priority document.

Respectfully submitted,



Howard L. Bernstein
Registration No. 25,665

SUGHRUE MION, PLLC
Telephone: (202) 293-7060
Facsimile: (202) 293-7860

WASHINGTON OFFICE

23373

CUSTOMER NUMBER

Enclosures: Japan 2003-275983

Date: August 12, 2004

【書類名】 特許願
【整理番号】 64003001
【特記事項】 特許法第 3 0 条第 1 項の規定の適用を受けようとする特許出願
【提出日】 平成15年 7月17日
【あて先】 特許庁長官殿
【国際特許分類】 G06N 7/00
【発明者】
 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内
 【氏名】 安東 正貴
【発明者】
 【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内
 【氏名】 斎藤 彰
【発明者】
 【住所又は居所】 広島県廿日市市宮園 9 丁目 1 の 7
 【氏名】 大瀧 慈
【発明者】
 【住所又は居所】 広島県広島市佐伯区楽々園 5 - 9 五日市住宅 1 6 - 2 0 2
 【氏名】 佐藤 健一
【発明者】
 【住所又は居所】 広島県広島市南区仁保南 2 丁目 8 - 7
 【氏名】 西山 正彦
【発明者】
 【住所又は居所】 東京都中央区八丁堀二丁目 2 6 番 9 号 グランデビルディング
 社団法人バイオ産業情報化コンソーシアム内
 【氏名】 大谷 敬子
【特許出願人】
 【識別番号】 000004237
 【氏名又は名称】 日本電気株式会社
【特許出願人】
 【識別番号】 503077165
 【氏名又は名称】 大瀧 慈
【特許出願人】
 【識別番号】 500535301
 【氏名又は名称】 社団法人バイオ産業情報化コンソーシアム
【代理人】
 【識別番号】 100071272
 【弁理士】
 【氏名又は名称】 後藤 洋介
【選任した代理人】
 【識別番号】 100077838
 【弁理士】
 【氏名又は名称】 池田 憲保
【手数料の表示】
 【予納台帳番号】 012416
 【納付金額】 21,000円
【その他】 国等の委託研究の成果に係る特許出願（平成 1 4 年度新エネルギー・産業技術総合開発機構「遺伝子多様性モデル解析事業」委託研究、産業活力再生特別措置法第 3 0 条の適用を受けるもの）
【提出物件の目録】
 【物件名】 特許請求の範囲 1



【物件名】 明細書 1
【物件名】 図面 1
【物件名】 要約書 1
【包括委任状番号】 0018587
【包括委任状番号】 0307208

【書類名】 特許請求の範囲

【請求項 1】

遺伝子発現強度データを送出する入力装置と、プログラム制御により動作するデータ解析装置と、出力装置とを含み、チャンネル毎に遺伝子が発現している確率を推定する遺伝子発現状況推定システムにおいて、

前記データ解析装置は、

前記入力装置から送出された遺伝子発現強度データを用いて、以下の数 1 に示される混合正規分布（尚、外 1 は平均 0、分散 σ^2 の 1 次元正規分布の密度関数を表し、外 2 - 1 及び外 2 - 2 はそれぞれ第 1 および第 2 のコンポーネントの平均と分散パラメータを表し、 ξ は混合率を表し、外 3 が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出する分布パラメータ推定手段と、

【数 1】

$$(1 - \xi) \phi(u - \mu_0 | \sigma_0^2) + \xi \phi(u - \mu_1 | \sigma_1^2)$$

[外 1]

$$\phi(* | \sigma^2)$$

[外 2 - 1]

$$(\mu_0, \sigma_0^2)$$

[外 2 - 2]

$$(\mu_1, \sigma_1^2)$$

[外 3]

$$\mu_0 < \mu_1, \sigma_0^2 > 0, \sigma_1^2 > 0, 0 < \xi < 1$$

前記入力装置から送出された遺伝子発現強度データと、前記分布パラメータ推定手段から送出された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出する混合比パラメータ推定手段と、

前記遺伝子発現強度データと、前記推定された分布パラメータ及び混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出する事後確率計算手段を有し、

計算された前記事後確率が前記出力装置に出力されることを特徴とする遺伝子発現状況推定システム。

【請求項 2】

前記分布パラメータ推定手段は、2 つのチャンネルの遺伝子発現強度 X、Y の差を示す V が 0 近傍である遺伝子に関する発現量の和のデータに対して、2 つのコンポーネントからなる前記混合正規分布を当てはめ、混合率（ ξ ）、平均（ μ_0 、 μ_1 ）、及び分散外 3 - 1 を推定する

[外 3 - 1]

$$(\sigma_0^2, \sigma_1^2)$$

ことを特徴とする請求項 1 記載の遺伝子発現状況推定システム。

【請求項 3】

前記遺伝子に関する発現量の和のデータは、遺伝子発現強度 X 、 Y の差の絶対値外 4 の中央値を c_M とするとき、外 5 で示される

[外 4]

$$|v_i| \quad (i = 1, \dots, n)$$

[外 5]

$$\{u_i \mid |v_i| \leq c_M, i = 1, \dots, n\}$$

ことを特徴とする請求項 2 記載の遺伝子発現状況推定システム。

【請求項 4】

前記分布パラメータ推定手段は、推定された外 6 を用いて、外 7 を以下の数 2、数 3（尚、 N_0 は外 8 を満たすデータのインデックス集合とし、外 9 はその要素の個数とする。）

、数 4、及び数 5 に従って推定する

【数 2】

$$\hat{\mu} = (\hat{\mu}_1 - \hat{\mu}_0) / 2$$

【数 3】

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2\|N_0\|} \sum_{i \in N_0} v_i^2$$

【数 4】

$$\hat{\sigma}_\beta^2 = \frac{1}{4} \hat{\sigma}_0^2 - \frac{1}{2} \hat{\sigma}_\varepsilon^2$$

【数 5】

$$\hat{\lambda} = \sqrt{\log \left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_0^2}{4\hat{\mu}^2} \right)}$$

[外 6]

$$\hat{\xi}, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2$$

[外 7]

$$\mu, \sigma_\varepsilon^2, \sigma_\beta^2, \lambda$$

[外 8]

$$i \in \{i \mid u_i < \hat{\mu}_0\}$$

[外 9]

$$\| N_0 \|$$

ことを特徴とする請求項 3 記載の遺伝子発現状況推定システム。

【請求項 5】

前記混合比パラメータ推定手段は、前記入力装置から送出された遺伝子発現強度データ $\{(u_i, v_i) \mid i = 1, \dots, n\}$ に対して、前記分布パラメータ推定手段から与えられた外 10 を用いて、以下の数 6（ただし、以下の数 7 及び数 8 に示される関係を満たすとする。）に示される 4 つのコンポーネントからなる 2 変量混合正規分布を当てはめ、混合比パラメータ $p = (p_{00}, p_{10}, p_{01}, p_{11})$ （ p_{00} は細胞 1, 2 において、共に遺伝子が発現していない場合における混合比パラメータを表し、 p_{11} は細胞 1, 2 において、共に遺伝子が発現している場合における混合比パラメータを表し、 p_{10} は細胞 1 において遺伝子が発現しており、細胞 2 において遺伝子が発現していない場合における混合比パラメータを表し、 p_{01} は細胞 1 において遺伝子が発現していなく、細胞 2 において遺伝子が発現している場合における混合比パラメータを表す）を推定する

[外 10]

$$\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2)$$

【数 6】

$$\begin{aligned} & p_{00}g_{00}(u, v \mid \hat{\theta}) + p_{10}g_{10}(u, v \mid \hat{\theta}) + p_{01}g_{01}(u, v \mid \hat{\theta}) + p_{11}g_{11}(u, v \mid \hat{\theta}) \\ & \equiv p_{00}\phi\left(u \mid 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) + p_{10}\phi_2(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10}) \\ & \quad + p_{01}\phi_2(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01}) + p_{11}\phi\left(u - 2\hat{\mu} \mid 4\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \right. \\ & \quad \left. + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) \end{aligned}$$

【数 7】

$$\hat{\Sigma}_{10} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_\varepsilon^2 \end{pmatrix}$$

【数 8】

$$\hat{\Sigma}_{01} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_{\varepsilon}^2 \end{pmatrix}$$

ことを特徴とする請求項 4 記載の遺伝子発現状況推定システム。

【請求項 6】

前記事後確率計算手段は、前記入力装置から送出された遺伝子発現強度データの対（ u ， v ）に対して、細胞 1 および細胞 2 における任意の遺伝子が発現している事後確率を、以下の数 9（尚、外 1 0 - 1 は、以下の数 1 0 によって与えられるものとする。）及び数 1 1（尚、 τ_1 ， τ_2 はそれぞれの細胞における遺伝子の真の発現の有無を 1 および 0 で表したものである。）にしたがって算出する

【数 9】

$$\Pr(\tau_1 = 1 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10}g_{10}(u, v | \hat{\theta}) + \hat{p}_{11}g_{11}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

【数 1 0】

$$f(u, v | \hat{p}, \hat{\theta}) = \sum_{(j,k) \in \{0,1\}^2} \hat{p}_{jk}g_{jk}(u, v | \hat{\theta})$$

【数 1 1】

$$\Pr(\tau_2 = 1 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{01}g_{01}(u, v | \hat{\theta}) + \hat{p}_{11}g_{11}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

[外 1 0 - 1]

$$f(u, v | \hat{p}, \hat{\theta})$$

ことを特徴とする請求項 5 記載の遺伝子発現状況推定システム。

【請求項 7】

前記事後確率計算手段は、細胞 1 および細胞 2 で遺伝子の発現状態が異なる状態である事後確率を、以下の数 1 2（尚、 τ_1 ， τ_2 はそれぞれの細胞における遺伝子の真の発現の有無を 1 および 0 で表したものである。）にしたがって算出する

【数 1 2】

$$\Pr(\tau_1 \neq \tau_2 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10}g_{10}(u, v | \hat{\theta}) + \hat{p}_{01}g_{01}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

ことを特徴とする請求項 5 記載の遺伝子発現状況推定システム。

【請求項 8】

前記事後確率計算手段は、全ての遺伝子発現強度データの対（ u ， v ）に基づいて、遺

伝子が発現している事後確率を計算したかどうかを判定し、計算していれば終了し、計算していなければ、次の遺伝子に関する事後確率を計算し、

計算された各チャンネル毎の遺伝子が発現している事後確率は、前記出力装置へ送出され、

該出力装置は、各チャンネル毎の遺伝子が発現している事後確率を表示することを特徴とする請求項6又は7記載の遺伝子発現状況推定システム。

【請求項9】

遺伝子発現強度データに基づいてチャンネル毎に遺伝子が発現している確率を推定する遺伝子発現状況推定方法において、

前記遺伝子発現強度データを用いて、以下の数13に示される混合正規分布（尚、外11は平均0、分散 σ^2 の1次元正規分布の密度関数を表し、外12-1及び外12-2はそれぞれ第1および第2のコンポーネントの平均と分散パラメータを表し、 ξ は混合率を表し、外13が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出するステップと、

【数13】

$$(1 - \xi) \phi(u - \mu_0 | \sigma_0^2) + \xi \phi(u - \mu_1 | \sigma_1^2)$$

[外11]

$$\phi(* | \sigma^2)$$

[外12-1]

$$(\mu_0, \sigma_0^2)$$

[外12-2]

$$(\mu_1, \sigma_1^2)$$

[外13]

$$\mu_0 < \mu_1, \sigma_0^2 > 0, \sigma_1^2 > 0, 0 < \xi < 1$$

前記遺伝子発現強度データと、前記推定された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出するステップと、

前記遺伝子発現強度データと、前記推定された分布パラメータ及び前記推定された混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出するステップと、

計算された前記事後確率を出力するステップ

を有することを特徴とする遺伝子発現状況推定方法。

【請求項10】

前記分布パラメータを推定するステップはさらに、2つのチャンネルの遺伝子発現強度X、Yの差を示すVが0近傍である遺伝子発現強度データに対して、2つのコンポーネントからなる前記混合正規分布を当てはめ、混合率（ ξ ）、平均（ μ_0 、 μ_1 ）、及び分散外13-1を推定するステップを有する

[外13-1]

$$(\sigma_0^2, \sigma_1^2)$$

ことを特徴とする請求項 9 記載の遺伝子発現状況推定方法。

【請求項 1 1】

前記遺伝子に関する発現量の和のデータは、遺伝子発現強度 X , Y の差の絶対値外 1 4 - 1 の中央値を c_M とするとき、外 1 4 - 2 で示される
[外 1 4 - 1]

$$|v_i| (i = 1, \dots, n)$$

[外 1 4 - 2]

$$\{u_i || |v_i| \leq c_M, i = 1, \dots, n\}$$

ことを特徴とする請求項 1 0 記載の遺伝子発現状況推定方法。

【請求項 1 2】

前記分布パラメータを推定するステップにおいて、推定された外 1 5 を用いて、外 1 6 を以下の数 1 4、数 1 5 (尚、 N_0 は外 1 7 を満たすデータのインデックス集合とし、外 1 8 はその要素の個数とする。)、数 1 6、及び数 1 7 に従って推定する

【数 1 4】

$$\hat{\mu} = (\hat{\mu}_1 - \hat{\mu}_0) / 2$$

【数 1 5】

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2 \|N_0\|} \sum_{i \in N_0} v_i^2$$

【数 1 6】

$$\hat{\sigma}_\beta^2 = \frac{1}{4} \hat{\sigma}_0^2 - \frac{1}{2} \hat{\sigma}_\varepsilon^2$$

【数 1 7】

$$\hat{\lambda} = \sqrt{\log \left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_0^2}{4 \hat{\mu}^2} \right)}$$

[外 1 5]

$$\xi, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2$$

[外 1 6]

$$\mu, \sigma_{\varepsilon}^2, \sigma_{\beta}^2, \lambda$$

[外 1 7]

$$i \in \{i \mid u_i < \hat{\mu}_0\}$$

[外 1 8]

$$\| N_0 \|$$

ことを特徴とする請求項 1 1 記載の遺伝子発現状況推定方法。

【請求項 1 3】

前記混合比パラメータを推定するステップにおいて、前記送出された遺伝子発現強度データ $\{(u_i, v_i) \mid i = 1, \dots, n\}$ に対して、前記送出された外 1 9 を用いて、以下の数 1 8 (ただし、以下の数 1 9 及び数 2 0 に示される関係を満たすとする。)に示される 4 つのコンポーネントからなる 2 変量混合正規分布を当てはめ、混合比パラメータ $p = (p_{00}, p_{10}, p_{01}, p_{11})$ (p_{00} は細胞 1, 2 において、共に遺伝子が発現していない場合における混合比パラメータを表し、 p_{11} は細胞 1, 2 において、共に遺伝子が発現している場合における混合比パラメータを表し、 p_{10} は細胞 1 において遺伝子が発現しており、細胞 2 において遺伝子が発現していない場合における混合比パラメータを表し、 p_{01} は細胞 1 において遺伝子が発現していなく、細胞 2 において遺伝子が発現している場合における混合比パラメータを表す) を推定する

【数 1 8】

$$\begin{aligned} & p_{00}g_{00}(u, v \mid \hat{\theta}) + p_{10}g_{10}(u, v \mid \hat{\theta}) + p_{01}g_{01}(u, v \mid \hat{\theta}) + p_{11}g_{11}(u, v \mid \hat{\theta}) \\ &= p_{00}\phi\left(u \mid 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2\right)\phi\left(v \mid 2\hat{\sigma}_{\varepsilon}^2\right) + p_{10}\phi_2(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10}) \\ &+ p_{01}\phi_2(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01}) + p_{11}\phi\left(u - 2\hat{\mu} \mid 4\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \right. \\ &\quad \left. + 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2\right)\phi\left(v \mid 2\hat{\sigma}_{\varepsilon}^2\right) \end{aligned}$$

【数 1 9】

$$\hat{\Sigma}_{10} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2 & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_{\varepsilon}^2 \end{pmatrix}$$

【数 2 0】

$$\hat{\Sigma}_{01} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_{\varepsilon}^2 \end{pmatrix}$$

[外 1 9]

$$\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_{\varepsilon}^2, \hat{\sigma}_{\beta}^2)$$

ことを特徴とする請求項 1 2 記載の遺伝子発現状況推定方法。

【請求項 1 4】

前記事後確率を計算するステップでは、前記送出された遺伝子発現強度データの対 (u, v) に対して、細胞 1 および細胞 2 における任意の遺伝子が発現している事後確率を、以下の数 2 1 (尚、外 1 9 - 1 は、以下の数 2 2 によって与えられるものとする。) 及び数 2 3 (尚、 τ_1 , τ_2 はそれぞれの細胞における遺伝子の真の発現の有無を 1 および 0 で表したものである。) にしたがって算出する

【数 2 1】

$$\Pr(\tau_1 = 1 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10}g_{10}(u, v | \hat{\theta}) + \hat{p}_{11}g_{11}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

【数 2 2】

$$f(u, v | \hat{p}, \hat{\theta}) = \sum_{(j,k) \in \{0,1\}^2} \hat{p}_{jk}g_{jk}(u, v | \hat{\theta})$$

【数 2 3】

$$\Pr(\tau_2 = 1 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{01}g_{01}(u, v | \hat{\theta}) + \hat{p}_{11}g_{11}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

[外 1 9 - 1]

$$f(u, v | \hat{p}, \hat{\theta})$$

ことを特徴とする請求項 1 3 記載の遺伝子発現状況推定方法。

【請求項 1 5】

前記事後確率を計算するステップでは、細胞 1 および細胞 2 で遺伝子の発現状態が異なる状態である事後確率を、以下の数 2 4 (尚、 τ_1 , τ_2 はそれぞれの細胞における遺伝子の真の発現の有無を 1 および 0 で表したものである。) にしたがって算出する

【数 2 4】

$$\Pr(\tau_1 \neq \tau_2 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10}g_{10}(u, v | \hat{\theta}) + \hat{p}_{01}g_{01}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

ことを特徴とする請求項 1 3 記載の遺伝子発現状況推定方法。

【請求項 1 6】

前記事後確率を計算するステップでは、全ての遺伝子発現強度データの対 (u, v) に基づいて、遺伝子が発現している事後確率を計算したかどうかを判定し、計算していれば終了し、計算していなければ、次の遺伝子に関する事後確率を計算する

ことを特徴とする請求項 1 4 又は 1 5 記載の遺伝子発現状況推定方法。

【請求項 1 7】

遺伝子発現強度データに基づいてチャンネル毎に遺伝子が発現している確率を推定するためコンピュータに、

前記遺伝子発現強度データを用いて、以下の数 2 5 に示される混合正規分布（尚、外 2 0 は平均 0、分散 σ^2 の 1 次元正規分布の密度関数を表し、外 2 1 - 1 及び外 2 1 - 2 はそれぞれ第 1 および第 2 のコンポーネントの平均と分散パラメータを表し、 ξ は混合率を表し、外 2 2 が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出するステップと、

【数 2 5】

$$(1 - \xi)\phi(u - \mu_0 | \sigma_0^2) + \xi\phi(u - \mu_1 | \sigma_1^2)$$

[外 2 0]

$$\phi(* | \sigma^2)$$

[外 2 1 - 1]

$$(\mu_0, \sigma_0^2)$$

[外 2 1 - 2]

$$(\mu_1, \sigma_1^2)$$

[外 2 2]

$$\mu_0 < \mu_1, \sigma_0^2 > 0, \sigma_1^2 > 0, 0 < \xi < 1$$

前記遺伝子発現強度データと、前記推定された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出するステップと、

前記遺伝子発現強度データと、前記推定された分布パラメータ及び前記推定された混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出するステップと、

計算された前記事後確率を出力するステップ

を実行させるための遺伝子発現状況推定プログラム。

【請求項 1 8】

遺伝子発現強度データに基づいてチャンネル毎に遺伝子が発現している確率を推定するためコンピュータに、

前記遺伝子発現強度データを用いて、以下の数 2 6 に示される混合正規分布（尚、外 2 3 は平均 0, 分散 σ^2 の 1 次元正規分布の密度関数を表し、外 2 4 - 1 及び外 2 4 - 2 はそれぞれ第 1 および第 2 のコンポーネントの平均と分散パラメータを表し、 ξ は混合率を表し、外 2 5 が満たされているものとする。）の分布パラメータを推定し、この推定された分布パラメータを送出するステップと、

【数 2 6】

$$(1 - \xi) \phi(u - \mu_0 | \sigma_0^2) + \xi \phi(u - \mu_1 | \sigma_1^2)$$

[外 2 3]

$$\phi(* | \sigma^2)$$

[外 2 4 - 1]

$$(\mu_0, \sigma_0^2)$$

[外 2 4 - 2]

$$(\mu_1, \sigma_1^2)$$

[外 2 5]

$$\mu_0 < \mu_1, \sigma_0^2 > 0, \sigma_1^2 > 0, 0 < \xi < 1$$

前記遺伝子発現強度データと、前記推定された分布パラメータとを用いて、混合正規分布の混合比パラメータを推定し、この推定された混合比パラメータを送出するステップと、

前記遺伝子発現強度データと、前記推定された分布パラメータ及び前記推定された混合比パラメータを用いて、各チャンネル毎に各遺伝子の発現状況に関する事後確率を計算し、計算された事後確率を送出するステップと、

計算された前記事後確率を出力するステップ

を実行させるための遺伝子発現状況推定プログラムを記録したコンピュータ読み取り可能な記録媒体。

【書類名】明細書

【発明の名称】遺伝子発現状況推定システム、方法、プログラム、及び記録媒体

【技術分野】

【0001】

本発明は、二色蛍光法による cDNA マイクロアレイデータの統計的解析方法、解析システム、及び記録媒体に関し、特にチャンネル毎に遺伝子が発現している確率を推定するシステム、方法、プログラム、及び記録媒体に関するものである。

【背景技術】

【0002】

現在、ゲノム研究は個々の遺伝子についての構造解析から体系的な遺伝子の機能解析へと展開しつつある。機能未知の遺伝子や総体としての遺伝子の機能解析のために、多数の遺伝子の発現強度を同時に定量化することのできる cDNA（相補的な DNA）マイクロアレイを用いた実験はその有効性が大いに期待されている。

【0003】

二色蛍光法による cDNA マイクロアレイを用いた実験の目的は二種類の細胞の遺伝子発現強度の違いを検出することにある。ここで、二色蛍光法による cDNA マイクロアレイの概要について述べる。まず、多数の遺伝子セットの cDNA を参照用のプローブとして、スライドガラス上にアレイ状に高密度に固定化する（マイクロアレイ）。

【0004】

次に、条件の異なる 2 種類のサンプル、細胞 1 と細胞 2（例えば正常細胞と癌細胞）から抽出した mRNA をそれぞれ波長の異なる蛍光色素でラベルし、ターゲット cDNA を合成する。そして、それらを等量混合したものをマイクロアレイに固定化された参照用のプローブ cDNA に競合的にハイブリダイズさせる。ハイブリダイゼーション後、スキャナーでそれぞれの蛍光色素強度を測定する。細胞 1 にラベルされた蛍光色素をチャンネル 1 により、細胞 2 にラベルされた蛍光色素をチャンネル 2 により読み取り、それぞれを各細胞の遺伝子発現強度データ（マイクロアレイデータ）とする。

【0005】

このように、マイクロアレイデータが得られるまでの過程は複雑であり、高度な実験技術が必要とされることから、実験の各段階において様々な実験誤差が生じると考えられる。このため、マイクロアレイデータから真に生物学的意味のあるデータを取り出すためには発現強度の分布と実験誤差の解析は解決すべき重要な課題である。

【0006】

発現強度の分布に関しては、例えば、以下の非特許文献 1 を参照すると、Newton 等は発現強度にガンマ分布関数を仮定し、発現強度比（チャンネル 1 とチャンネル 2 の発現強度の比）についての統計学的性質を考察している。

【0007】

また、観測された発現強度データに対しては、例えば、以下の非特許文献 2 を参照すると、Lee 等は真の発現強度を 2 個の水準値に分離できることおよび偶然誤差の存在を前提として、以下の数 2 7 に示されるような混合正規分布を適用し、発現強度データについての統計学的考察を行っている。

【0008】

【数 2 7】

$$f(x) = p\phi(x - \mu_1 | \sigma_1^2) + (1 - p)\phi(x - \mu_2 | \sigma_2^2)$$

ここで、 x はスキャナーなどによって得られる蛍光強度などの遺伝子発現強度（の対数値）を表し、右辺第 1 項の外 2 6 は遺伝子が発現しているときの平均 μ_1 、分散外 2 7 の正規分布、

[外 2 6]

$$\phi(\mathbf{x} - \mu_1 | \sigma_1^2)$$

[外 2 7]

$$\sigma_1^2$$

また、同第 2 項の外 2 8 は遺伝子が発現していないときの平均 μ_2 、分散外 2 9 の正規分布の密度関数を表し、 p はその混合率を表す母数である。

【 0 0 0 9 】

[外 2 8]

$$\phi(\mathbf{x} - \mu_2 | \sigma_2^2)$$

[外 2 9]

$$\sigma_2^2$$

【 0 0 1 0 】

実験誤差の解析については、系統誤差の除去、いわゆるノーマライゼーションの方法がいくつか提案されている。非特許文献 3 を参照すると、Chen 等は二つの細胞の遺伝子発現強度の中央値は等しいとしてチャンネル 1 とチャンネル 2 で得られた測定値の補正を行っている。また、非特許文献 4、5、6 を参照すると、Dudoit や Schuchhardt や Yang は、系統誤差がスポットのスライドガラス上の位置や、二種類の蛍光色素の感度の違いによって生じたものと考え、それらを除去する方法を提案している。

【非特許文献 1】Newton et. al、2001 年、ジャーナル・オブ・コンピュテーショナル・バイオロジー、第 8 巻、37～52 頁 (Journal of Computational Biology Vol. 8, pp. 37-52)

【非特許文献 2】Lee et. al、2000 年、プロシーディング・オブ・ザ・ナショナル・アカデミー・オブ・サイエンス、第 97 巻、第 18 号、9834～9839 頁 (Proceeding of the National Academy of Sciences Vol. 97, No 18, pp. 9834-9839)

【非特許文献 3】Chen et. al、1997 年、ジャーナル・オブ・バイオメディカル・オプティクス、第 2 号、364～374 頁 (Journal of Biomedical Optics Vol. 2, pp. 364-374)

【非特許文献 4】Dudoit et. al、2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report #578 2.

【非特許文献 5】Schuchhardt et. al、2000 年、ヌクレ・アシッド・リサーチ、第 28 巻、第 10 号 (Nucleic Acids Research, 2000, Vol.28, No. 10)

【非特許文献 6】Yang et. al、2002 年、ヌクレ・アシッド・リサーチ、第 30 巻、第 4 号 (Nucleic Acids Research, 2002, Vol.30, No. 4)

【発明の開示】

【発明が解決しようとする課題】

【 0 0 1 1 】

上記した従来技術における問題点は、マイクロアレイデータの解析結果は再現性に乏しく解析の精度や効率が低いことにある。その原因として、従来の解析方法では、得られたマイクロアレイデータが真の信号と系統誤差および測定誤差に十分に分離できていないことが考えられる。したがって、系統誤差の除去および測定誤差の評価が重要な課題となる

。

【0012】

系統誤差の除去に関しては「cDNA マイクロアレイデータの補正法、システム、及び記録媒体」として、別途特許にて申請中である。本発明におけるマイクロアレイデータは、系統誤差はあらかじめ取り除かれているものとする。

【0013】

従来のマイクロアレイデータを用いた解析では、各スポットのcDNA量の均一性が保証されていないという理由で、2つのチャンネルの遺伝子発現強度の比（対数値の差）のみを扱っており、各チャンネル毎の遺伝子発現強度を扱っていない。従って、遺伝子の発現状況に関する真の信号と測定誤差の分離が十分に行われていない。

【0014】

本発明の目的は、マイクロアレイデータを用いた解析の精度および効率を高めるために、遺伝子の発現に関する真の信号と測定誤差の分離を行い、さらに、各チャンネル毎に、遺伝子が発現している確率を推定する手法およびシステムを提供することにある。

【課題を解決するための手段】**【0015】**

本発明の遺伝子発現状況推定システムは、マイクロアレイデータを入力する入力装置と、プログラム制御により動作するデータ解析装置と、出力装置とを含む。前記データ解析装置は、前記入力装置から与えられた遺伝子発現強度データを用いて、混合正規分布のコンポーネント毎の分布パラメータおよび混合比パラメータを推定するパラメータ推定手段と、推定された各パラメータを用いて、各チャンネル毎に遺伝子の発現に関する事後確率を計算する事後確率計算手段を有し、計算された事後確率は前記出力装置に出力することを特徴とする。

【0016】

このような構成を採用し、遺伝子発現状況を推定することにより、本発明の目的は達成することができる。

【発明の効果】**【0017】**

本発明による第1の効果は、マイクロアレイによって得られた遺伝子発現強度データに対して遺伝子の発現および非発現という概念を導入し、数理モデルを構築することにより遺伝子の発現に関する真の信号と実験誤差との分離が可能となったことにある。

【0018】

本発明による第2の効果は、マイクロアレイによって得られた遺伝子発現強度データを、2つのチャンネルの遺伝子発現強度データの和と差のデータに変換することにより、2つのチャンネルの蛍光強度の感度情報が得られ易くなったことにある。この結果、感度の違いによる実験誤差の大きさを視覚的に示すことが可能となった。

【0019】

本発明による第3の効果は、マイクロアレイによって得られた遺伝子発現強度データを、2つのチャンネルの遺伝子発現強度の和と差のデータに変換し、それらの2次元同時分布を記述することにより、2つのチャンネル毎に、各遺伝子の発現および非発現に関する事後確率を推定することが可能となったことにある。その結果、細胞1と細胞2の違いに関係する遺伝子を高い精度で検出することが可能となった。

【発明を実施するための最良の形態】**【0020】**

はじめに本発明におけるマイクロアレイによって得られた遺伝子発現強度データに対する数理モデルを説明する。Xをチャンネル1によって得られた細胞1の遺伝子発現強度とし、Yをチャンネル2によって得られた細胞2の遺伝子発現強度としたときに、それぞれの遺伝子発現データを数28で示す。なお、XおよびYは観察値に対して対数変換ないしは適当なべき変換および1次変換を含む適当な変換が施されている量である。

【0021】

【数 2 8】

$$X = \tau_1 \alpha + \beta + \varepsilon_1$$

$$Y = \tau_2 \alpha + \beta + \varepsilon_2$$

【0 0 2 2】

ここで、 τ_1 、 τ_2 はそれぞれの細胞における遺伝子の真の発現の有無（ON/OFF）を 1 および 0 で表す。また、 α は遺伝子が ON の状態にある場合に産生される mRNA の量およびスポットの状態に規定される遺伝子発現強度変量とし、 β はチャンネル 1 およびチャンネル 2 に共通な測定誤差とし、 ε はチャンネル間で独立な測定誤差とする。なお、それぞれの確率変数の分布は数 2 9 に従うものとする。

【0 0 2 3】

【数 2 9】

$$\log \alpha \sim N\left(\mu - \frac{\lambda^2}{2}, \lambda^2\right)$$

$$\varepsilon_j \sim N(0, \sigma_\varepsilon^2), j = 1, 2$$

$$\beta \sim N(0, \sigma_\beta^2)$$

【0 0 2 4】

ここで、 $N(\mu, \sigma^2)$ は平均 μ 、分散 σ^2 の 1 次元正規分布とする。また、 α 、 β および ε はそれぞれ独立とする。この数理モデルにおいては、遺伝子が発現している場合（ON 状態）の真の発現強度は非負値をとる確率変数であり、遺伝子が発現していない場合（OFF 状態）は、単なる実験誤差のみが観察されるとしている。さらに、非特許文献 6 を参照して、Yang YH らが導入した M-A 変換を変形した数 3 0 で示すような S-D 変換を行う。

【0 0 2 5】

【数 3 0】

$$U = X + Y,$$

$$V = X - Y,$$

【0 0 2 6】

すなわち、2 つのチャンネルの遺伝子発現強度の和を U とし、差を V と変換する。この S-D 変換を行ったときの本モデルの模式図を図 1 に示す。なお、以下ではこのプロットを S-D プロットとよぶ。図 1 において、 g_{00} はそれぞれの細胞において遺伝子が発現していない同時分布を表し、 g_{10} は細胞 1 において遺伝子が発現しており、細胞 2 において遺伝子が発現していない同時分布を表し、 g_{01} は細胞 1 において遺伝子が発現していない、細胞 2 において遺伝子が発現している同時分布を表し、 g_{11} はそれぞれの細胞において遺伝子が発現している同時分布を表している。 g_{00} の密度関数を数 3 1 に示す。

【0 0 2 7】

【数 3 1】

$$g_{00}(u, v | \theta) = \phi\left(u | 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2\right) \phi\left(v | 2\sigma_{\varepsilon}^2\right)$$

ここで、 $\phi(u | \sigma^2)$ は平均 0 分散 σ^2 の 1 次元正規分布の密度関数とする。 g_{10} の密度関数を数 3 2 に示す。

【0 0 2 8】

【数 3 2】

$$g_{10}(u, v | \theta)$$

$$= \int_{-\infty}^{+\infty} \phi\left(u - \mu e^{-\frac{\lambda^2}{2} + \lambda z} | 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2\right) \phi\left(v - \mu e^{-\frac{\lambda^2}{2} + \lambda z} | 2\sigma_{\varepsilon}^2\right) \phi(z | 1) dz$$

$$\equiv \phi_2(u - \mu, v - \mu | \Sigma_{10})$$

ここで、 $\phi_2(u, v | \Sigma)$ は平均ベクトル 0、分散共分散行列 Σ の 2 次元正規分布の密度関数とし、 Σ_{10} は 2×2 の分散共分散行列であり、数 3 3 に示す。

【0 0 2 9】

【数 3 3】

$$\Sigma_{10} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2 & \mu^2(e^{\lambda^2} - 1) \\ \mu^2(e^{\lambda^2} - 1) & \mu^2(e^{\lambda^2} - 1) + 2\sigma_{\varepsilon}^2 \end{pmatrix}$$

g_{01} の密度関数を数 3 4 に示す。

【0 0 3 0】

【数 3 4】

$$g_{01}(u, v | \theta)$$

$$= \int_{-\infty}^{+\infty} \phi\left(u - \mu e^{-\frac{\lambda^2}{2} + \lambda z} | 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2\right) \phi\left(v + \mu e^{-\frac{\lambda^2}{2} + \lambda z} | 2\sigma_{\varepsilon}^2\right) \phi(z | 1) dz$$

$$\equiv \phi_2(u - \mu, v + \mu | \Sigma_{01})$$

ここで、 Σ_{01} は 2×2 の分散共分散行列であり、数 3 5 に示す。

【0 0 3 1】

【数 3 5】

$$\Sigma_{01} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2 & -\mu^2(e^{\lambda^2} - 1) \\ -\mu^2(e^{\lambda^2} - 1) & \mu^2(e^{\lambda^2} - 1) + 2\sigma_{\varepsilon}^2 \end{pmatrix}$$

g_{11} の密度関数を数 3 6 に示す。

【0 0 3 2】

【数 3 6】

$$\begin{aligned} g_{11}(u, v | \theta) &= \phi(v | 2\sigma_{\varepsilon}^2) \int_{-\infty}^{+\infty} \phi\left(u - \mu e^{-\frac{\lambda^2}{2} + \lambda z} | 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2\right) \phi(z | 1) dz \\ &\equiv \phi\left(u - 2\mu | 4\mu^2(e^{\lambda^2} - 1) + 4\sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2\right) \phi(v | 2\sigma_{\varepsilon}^2) \end{aligned}$$

以上の分布をもとに、細胞 1 および細胞 2 における遺伝子が発現している事後確率を数 3 7 および数 3 8 に示す。

【0 0 3 3】

【数 3 7】

$$\Pr(\tau_1 = 1 | p, \theta) = \frac{p_{10}g_{10}(u, v | \theta) + p_{11}g_{11}(u, v | \theta)}{f(u, v | p, \theta)}$$

【数 3 8】

$$\Pr(\tau_2 = 1 | p, \theta) = \frac{p_{01}g_{01}(u, v | \theta) + p_{11}g_{11}(u, v | \theta)}{f(u, v | p, \theta)}$$

ここで、 $f(u, v | p, \theta)$ は数 3 9 によって与えられるものとする。

【0 0 3 4】

【数 3 9】

$$f(u, v | p, \theta) = \sum_{(j,k) \in \{0,1\}^2} p_{jk} g_{jk}(u, v | \theta)$$

ただし、ただし、 $p = (p_{00}, p_{10}, p_{01}, p_{11})$ は各分布の混合率を表すパラメータとする。

【0 0 3 5】

次に、本発明の実施の形態について図面を参照して詳細に説明する。図 2 を参照すると、本発明の第 1 の実施の形態は、遺伝子発現強度データに関する数理モデルの定式化およびそのデータ解析への適用による未知母数の推定する過程及び算出された母数の推定値を

使用して、細胞 1 と細胞 2 の遺伝子の発現状態に関する事後推定を行うためのシステムである。同システムにはキーボード等の入力装置 1 と、プログラムの制御により動作するデータ解析装置 2 と、ディスプレイ装置や印刷装置等の出力装置 3 が含まれる。

【0036】

データ解析装置 2 は、分布パラメータ推定手段 2 1 と、混合比パラメータ推定手段 2 2 と、事後確率計算手段 2 3 とを備えている。分布パラメータ推定手段 2 1 は、入力装置 1 から与えられた遺伝子発現強度データを用いて、混合正規分布の各コンポーネント毎の分布パラメータを推定する。推定された分布パラメータは、混合比パラメータ推定手段 2 2 と、事後確率計算手段 2 3 へ送られる。混合比パラメータ推定手段 2 2 は、入力装置 1 から与えられた遺伝子発現強度データと、分布パラメータ推定手段 2 1 から与えられた各コンポーネント毎の分布パラメータを用いて、混合正規分布の混合比パラメータを条件付き最尤法により推定する。推定された混合比パラメータは、事後確率計算手段 2 3 へ送られる。事後確率計算手段 2 3 は、入力装置 1 から与えられた遺伝子発現強度データと、分布パラメータ推定手段 2 1 から与えられた各コンポーネント毎の分布パラメータと、混合比パラメータ推定手段 2 2 から与えられた混合比パラメータを用いて、各チャンネル毎に遺伝子の発現状況に関する事後確率を計算する。計算された事後確率は出力装置 3 へ送られる。

【0037】

次に、図 2 及び図 3 を参照して、遺伝子発現強度データに関する数理モデルの定式化およびそのデータ解析への適用による未知母数の推定する過程について詳細に説明する。入力装置 1 から与えられた遺伝子発現強度データ外 3 0 は、分布パラメータ推定手段 2 1 及び混合比パラメータ推定手段 2 2 へ送られる。分布パラメータ推定手段 2 1 は遺伝子発現強度の差の絶対値外 3 1 の中央値を c_M とするとき、 $V=0$ 近傍の遺伝子に関する発現量の和のデータ外 3 2 に対して、以下の数 4 0 に示す、2 つのコンポーネントからなる混合正規分布を当てはめ、 ξ , μ_0 , σ_0 , μ_1 , σ_1 を推定する（図 3 のステップ A 1）。

【0038】

【数 4 0】

$$(1 - \xi) \phi(u - \mu_0 | \sigma_0^2) + \xi \phi(u - \mu_1 | \sigma_1^2)$$

【外 3 0】

$$\{(u_i, v_i) | i = 1, \dots, n\}$$

【外 3 1】

$$|v_i| (i = 1, \dots, n)$$

【外 3 2】

$$\{u_i | |v_i| \leq c_M, i = 1, \dots, n\}$$

【0039】

ここで、外 3 3 は、平均 0 分散 σ^2 の 1 次元正規分布の密度関数であり、外 3 4 と外 3 5 はそれぞれ第 1 および第 2 コンポーネントの平均と分散パラメータ、 ξ は混合率であり、外 3 6 が満たされているものとする。

【0040】

次に、分布パラメータ推定手段 2 1 は、推定された外 3 7 を用いて、外 3 8 を以下の数

4 1、数 4 2、数 4 3、及び数 4 4 に従って推定する（ステップ A 2）。

【 0 0 4 1 】

[外 3 3]

$$\phi \left(* \mid \sigma^2 \right)$$

[外 3 4]

$$\left(\mu_0, \sigma_0^2 \right)$$

[外 3 5]

$$\left(\mu_1, \sigma_1^2 \right)$$

[外 3 6]

$$\mu_0 < \mu_1, \sigma_0^2 > 0, \sigma_1^2 > 0, 0 < \xi < 1$$

[外 3 7]

$$\hat{\xi}, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2$$

[外 3 8]

$$\mu, \sigma_{\varepsilon}^2, \sigma_{\beta}^2, \lambda$$

【数 4 1】

$$\hat{\mu} = (\hat{\mu}_1 - \hat{\mu}_0) / 2$$

【数 4 2】

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{2 \| N_0 \|} \sum_{i \in N_0} v_i^2$$

【数 4 3】

$$\hat{\sigma}_{\beta}^2 = \frac{1}{4} \hat{\sigma}_0^2 - \frac{1}{2} \hat{\sigma}_{\varepsilon}^2$$

【数 4 4】

$$\hat{\lambda} = \sqrt{\log \left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_0^2}{4\hat{\mu}^2} \right)}$$

【0 0 4 2】

ここで、 N_0 は外 3 9 を満たすデータのインデックス集合とし、外 4 0 はその要素の個数とする。

【0 0 4 3】

[外 3 9]

$$i \in \{i \mid u_i < \hat{\mu}_0\}$$

[外 4 0]

|| N_0 ||

【0 0 4 4】

次に、混合比パラメータ推定手段 2 2 は、入力装置 1 から与えられた遺伝子発現強度データ $\{(u_i, v_i) \mid i = 1, \dots, n\}$ に対して、分布パラメータ推定手段 2 1 から与えられた各パラメータの推定値外 4 1 を用いて、以下の数 4 5 (尚、以下の数 4 6 (ただし、外 4 2 は数 3 3 から導かれる 2×2 の分散共分散行列) 及び数 4 7 (ただし、外 4 3 は数 3 5 から導かれる 2×2 の分散共分散行列) に示される関係を満たすとする。) に示される 4 つのコンポーネントからなる 2 変量混合正規分布を当てはめ、混合比パラメータ $p = (p_{00}, p_{10}, p_{01}, p_{11})$ を条件付き最尤法により推定する (ステップ A 3)。

【0 0 4 5】

[外 4 1]

$$\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2)$$

【数 4 5】

$$\begin{aligned} & p_{00}g_{00}(u, v \mid \hat{\theta}) + p_{10}g_{10}(u, v \mid \hat{\theta}) + p_{01}g_{01}(u, v \mid \hat{\theta}) + p_{11}g_{11}(u, v \mid \hat{\theta}) \\ &= p_{00}\phi\left(u \mid 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) + p_{10}\phi_2(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10}) \\ &+ p_{01}\phi_2(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01}) + p_{11}\phi\left(u - 2\hat{\mu} \mid 4\hat{\mu}^2(e^{\hat{\lambda}^2} - 1)\right. \\ &\left.+ 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2\right)\phi\left(v \mid 2\hat{\sigma}_\varepsilon^2\right) \end{aligned}$$

【数 4 6】

$$\hat{\Sigma}_{10} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2 & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_{\varepsilon}^2 \end{pmatrix}$$

【数 4 7】

$$\hat{\Sigma}_{01} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_{\beta}^2 + 2\hat{\sigma}_{\varepsilon}^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_{\varepsilon}^2 \end{pmatrix}$$

[外 4 2]

 $\hat{\Sigma}_{10}$

[外 4 3]

 $\hat{\Sigma}_{01}$

【0 0 4 6】

次に、算出された母数の推定値を使用して、細胞 1 と細胞 2 の遺伝子の発現状態に関する事後推定を行う過程を説明する。

【0 0 4 7】

各細胞の遺伝子毎の発現状況に関する事後確率計算手段 2 3 は、入力装置 1 から与えられた遺伝子発現強度データの対 (u, v) に対して、分布パラメータ推定手段 2 1 及び混合比パラメータ推定手段 2 2 から与えられた各パラメータの推定値外 4 4 および外 4 5 を用いて記述できる。

【0 0 4 8】

[外 4 4]

$$\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_{\varepsilon}^2, \hat{\sigma}_{\beta}^2)$$

[外 4 5]

$$\hat{P} = (\hat{P}_{00}, \hat{P}_{10}, \hat{P}_{01}, \hat{P}_{11})$$

【0 0 4 9】

すなわち、細胞 1 および細胞 2 における任意の遺伝子の発現が ON になっている事後確率は、以下の数 4 8 及び数 4 9 より算出できる (ステップ A 4)。

【0 0 5 0】

【数 4 8】

$$\Pr(\tau_1 = 1 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10}g_{10}(u, v | \hat{\theta}) + \hat{p}_{11}g_{11}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

【数 4 9】

$$\Pr(\tau_2 = 1 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{01}g_{01}(u, v | \hat{\theta}) + \hat{p}_{11}g_{11}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

【0 0 5 1】

全ての遺伝子発現強度データの対 (u, v) に基づいて、遺伝子の発現が ON になっている事後確率を計算したかどうかを判定し (ステップ A 5)、計算していれば終了する。計算していなければ、次の遺伝子に関する事後確率を計算する。

【0 0 5 2】

計算された各チャンネル毎の遺伝子が発現している事後確率は、出力装置 3 へ送る。出力装置 3 は、各チャンネル毎の遺伝子が発現している事後確率をグラフで表示したり、印刷したりする。

【0 0 5 3】

次に、本実施の形態の効果について説明する。本実施の形態では、マイクロアレイで得られた遺伝子発現データに対して、遺伝子の発現および非発現を導入した数理モデルを構築し、真の信号と実験誤差との分離を行った。また、2つのチャンネルの遺伝子発現強度による和と差のデータを用いることにより、マイクロアレイデータでの各チャンネルの蛍光強度の感度情報が得られ易くなり、実験誤差の大きさをよりの確に抽出することが可能となった。さらに、これらの和と差のデータに対して2次元同時分布を記述することにより、チャンネル毎に各遺伝子の発現に関して、高い精度の事後確率を推定することが可能となった。

【0 0 5 4】

また、細胞 1 と細胞 2 で遺伝子の発現状態が異なるという事象 (ON-OFF が不一致である状態) (図 3 のステップ A 4) の事後確率は以下の数 5 0 により算出される。

【0 0 5 5】

【数 5 0】

$$\Pr(\tau_1 \neq \tau_2 | \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10}g_{10}(u, v | \hat{\theta}) + \hat{p}_{01}g_{01}(u, v | \hat{\theta})}{f(u, v | \hat{p}, \hat{\theta})}$$

【0 0 5 6】

本実施の形態の効果として、細胞 1 と細胞 2 において発現状態が異なる可能性の高い遺伝子の候補を検出することが可能となる。

【0 0 5 7】

次に、本発明の第 2 の実施の形態について図面を参照して詳細に説明する。

【0 0 5 8】

図 4 を参照すると、本発明の第 2 の実施形態は、本発明の第 1 の実施形態と同様に、入力装置、データ解析装置、出力装置を備え、更に、データ解析プログラムを記録した記録媒体 4 を備える。この記録媒体 4 は可搬形あるいは固定型のいずれであってもよく、磁気ディスク、半導体メモリ、CD-ROM その他の記録媒体であってもよい。また、本手法を実行できるコンピュータプログラムを、ネットワークに接続されたコンピュータの記録装置に格納しておき、ネットワークを介して他のコンピュータに転送することもできる。本アルゴリズムを実行するコンピュータプログラムを提供する提供媒体としては、様々な形

式のコンピュータに読み出し可能な媒体として頒布可能であって、特定のタイプの媒体に限定されるものではない。

【 0 0 5 9 】

データ解析プログラムは記録媒体 4 からデータ解析装置 5 に読み込まれ、データ解析装置 5 の動作を制御し、入力装置 1 から入力されたデータファイルに対して第 1 の実施の形態におけるデータ処理装置 2 による処理と同一の処理を実行する。

【実施例 1】

【 0 0 6 0 】

次に、本発明の実施例について説明する。例として用いたデータは、異なる 2 種類の癌細胞（細胞 1、細胞 2）の遺伝子発現状況の比較のために行われた実験から得られたものである。

【 0 0 6 1 】

一枚のチップ上に 4 8 グリッド、1 グリッドあたり 4 4 1 (2 1 × 2 1) スポット、計 2 1 1 6 8 の遺伝子の発現パターンについて調べたものである。

【 0 0 6 2 】

図 5 および図 6 に、細胞 1 および細胞 2 の遺伝子の発現状態が両者ともに OFF、あるいは両者ともに ON である場合 ($V = 0$)、それらの発現強度の和の分布 U に混合正規分布と、その対比として単峰の正規分布を当てはめた場合の各分布の推定結果を示す。以下の表 1 にコンポーネント毎の分布パラメータの推定結果を示す（図 3 のステップ A 1 の結果）。

【 0 0 6 3 】

【表 1】

Result of N2MIXFit (Ver 0.998)

```
=====
Name of Data Set to be analyzed = n145h1.std
Name of Target Variate = S_CH1
Type of Transformation = 1/16
Sample size = 14726
Critical Value for Convergence = .10000E+06
Iterations for Convergence = 50
Job Termination Status = Normally Terminated
=====
```

	Mean	SD	Rate(%)
Single Component:	3.0273	4.3560	100.00
1st Component:	.89956	1.2978	47.38
2nd Component:	4.9430	5.1395	52.62

```
-----
Log_Likelihood for Single Component = -42565.
for Two Components Mixed = -40465.
Log of Likelihood Ratio Statistics = 2099.8
-----
```

【 0 0 6 4 】

図 5 に推定された累積分布関数を、図 6 に推定された密度関数を示す。細実線は混合正規分布を仮定した場合、二点鎖線はその第 1 コンポーネント (OFF-OFF)、太実線は第 2 コンポーネント (ON-ON)、そして点線は単峰の正規分布を仮定した場合の推定結果を示している。

【 0 0 6 5 】

図 5 における一点鎖線は観察されたデータに基づく経験累積分布関数を示しており、観察値が推定された混合正規分布（細実線）によく従っていることを示している。

【 0 0 6 6 】

図 6 において * 印（尚、左右両端は * 印が確認できるが、遺伝子発現強度が 0 付近から

3 0 付近までは重なり合っているため*印ではなく以下の(1)～(5)に示すハッチングで表示することとした。)は観察データを示しており、それぞれのハッチング領域((1)～(5))は、第1コンポーネントに属している事後確率の大きさを示している。(1)の黒塗り領域は0～0.2、(2)のハッチング領域は0.2～0.4、(3)のハッチング領域は0.4～0.6、(4)のハッチング領域は0.6～0.8、(5)のハッチング領域は0.8～1.0を表している。

【0067】

図7に、遺伝子発現強度データのS-Dプロットを示す。横軸は細胞1と細胞2の遺伝子発現強度の対数値の和を、縦軸はそれらの差を示している。各マークの色は、細胞1と細胞2の遺伝子の発現状態が不一致(ON-OFFあるいはOFF-ON)である事後確率の大きさを示している。(1)の黒塗り領域は0～0.2、(2)のハッチング領域は0.2～0.4、(3)のハッチング領域は0.4～0.6、(4)のハッチング領域は0.6～0.8、(5)のハッチング領域は0.8～1.0を表している。

【0068】

以下の表2に、条件付き最尤法による混合比パラメータの推定結果を示す(図3のステップA2, A3の結果)。図7における各プロットに対しては、それぞれ遺伝子に対応しているので、細胞1と細胞2の違いに関係する遺伝子の候補を容易に絞り込むことができる。

【0069】

【表2】

RESULT OF MAD Ver(0.998)

=====	
Transformation = 6	(1/16)
Samplre Size =	21168

SIGMA_Epsilon =	.898
MU =	2.022
LAMBDA =	.960
SIGMA_Beta =	.133
P11 =	.561
P10 =	.004
P01 =	.011
P00 =	.425

【図面の簡単な説明】

【0070】

【図1】本発明における数理モデルのS-Dプロットによる模式図である。

【図2】本発明の第1の実施の形態の構成を示すブロック図である。

【図3】本発明の第1の実施の形態の動作を示すフローチャートである。

【図4】本発明の第2の実施形態の構成を示すブロック図である。

【図5】V=0近傍の遺伝子発現強度データに対して推定した正規分布の累積分布グラフである。

【図6】V=0近傍の遺伝子発現強度データに対して推定した正規分布の密度関数を示したグラフである。

【図7】遺伝子発現強度データのS-Dプロットを示した図である。

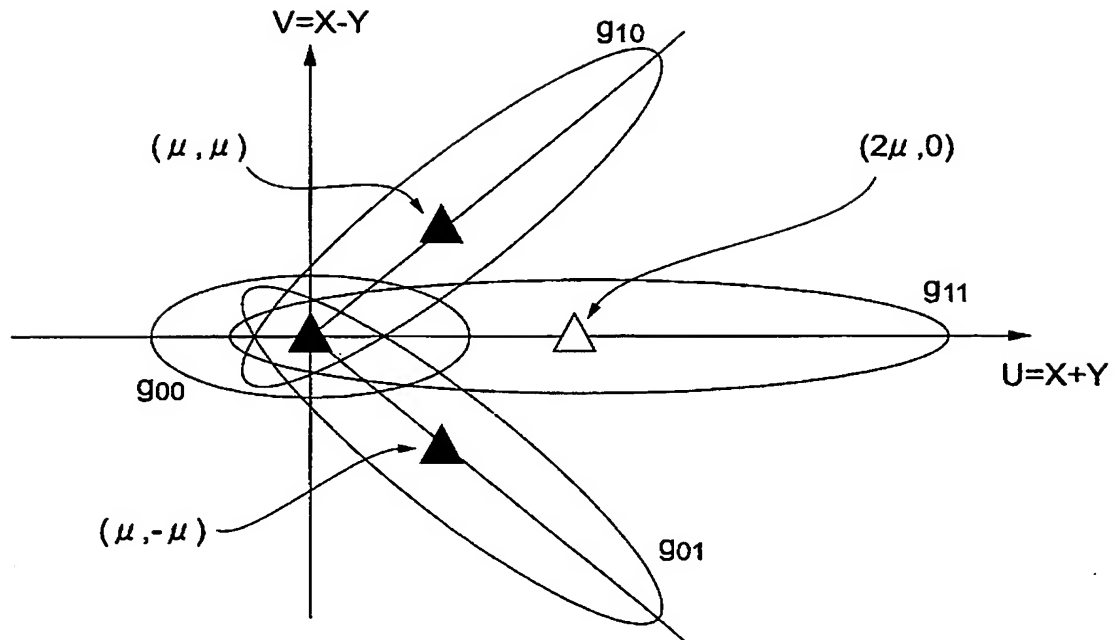
【符号の説明】

【 0 0 7 1 】

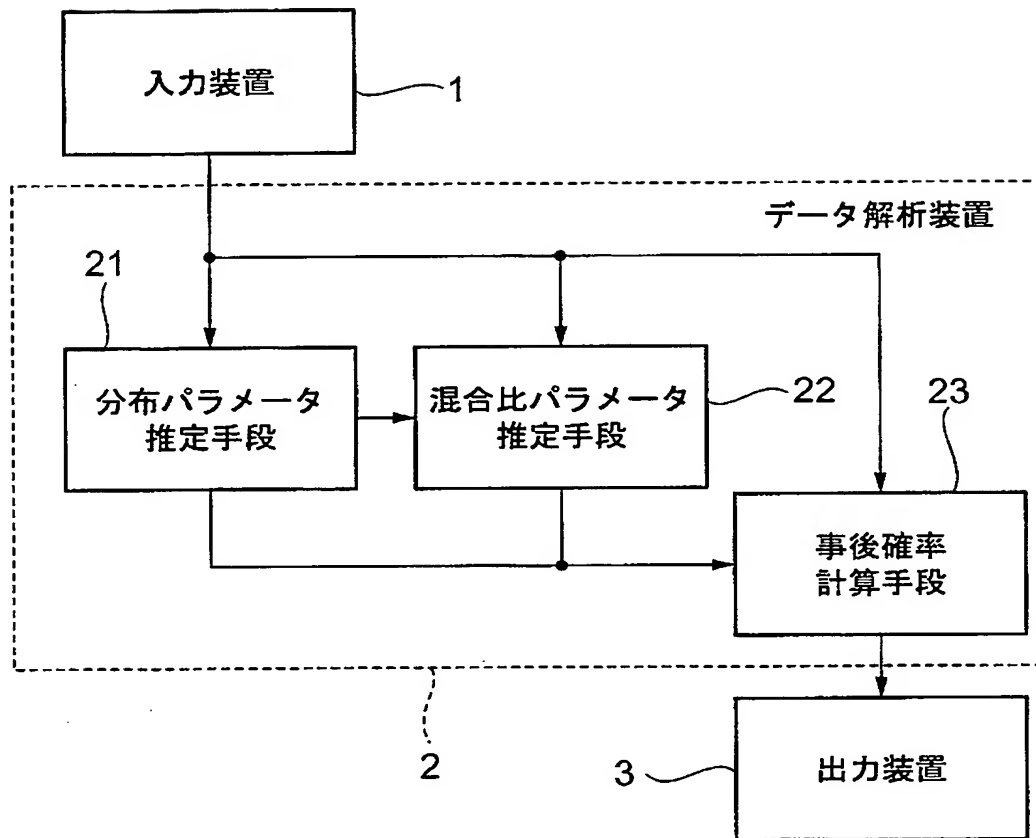
- 1 入力装置
- 2 データ解析装置
- 3 出力装置
- 4 記録媒体
- 5 データ解析装置
- 2 1 分布パラメータ推定手段
- 2 2 混合比パラメータ推定手段
- 2 3 事後確率計算手段

【書類名】 図面

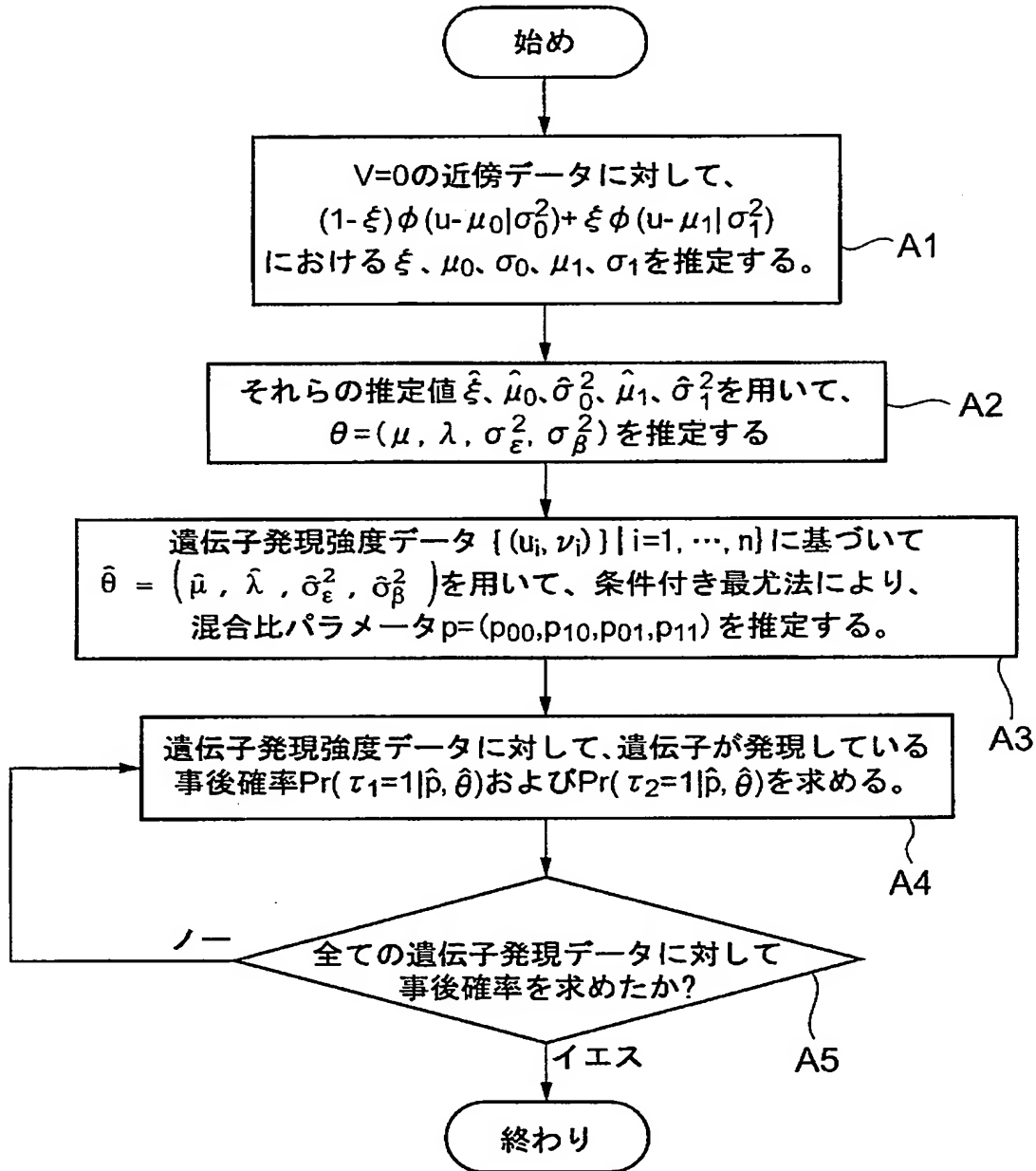
【図 1】



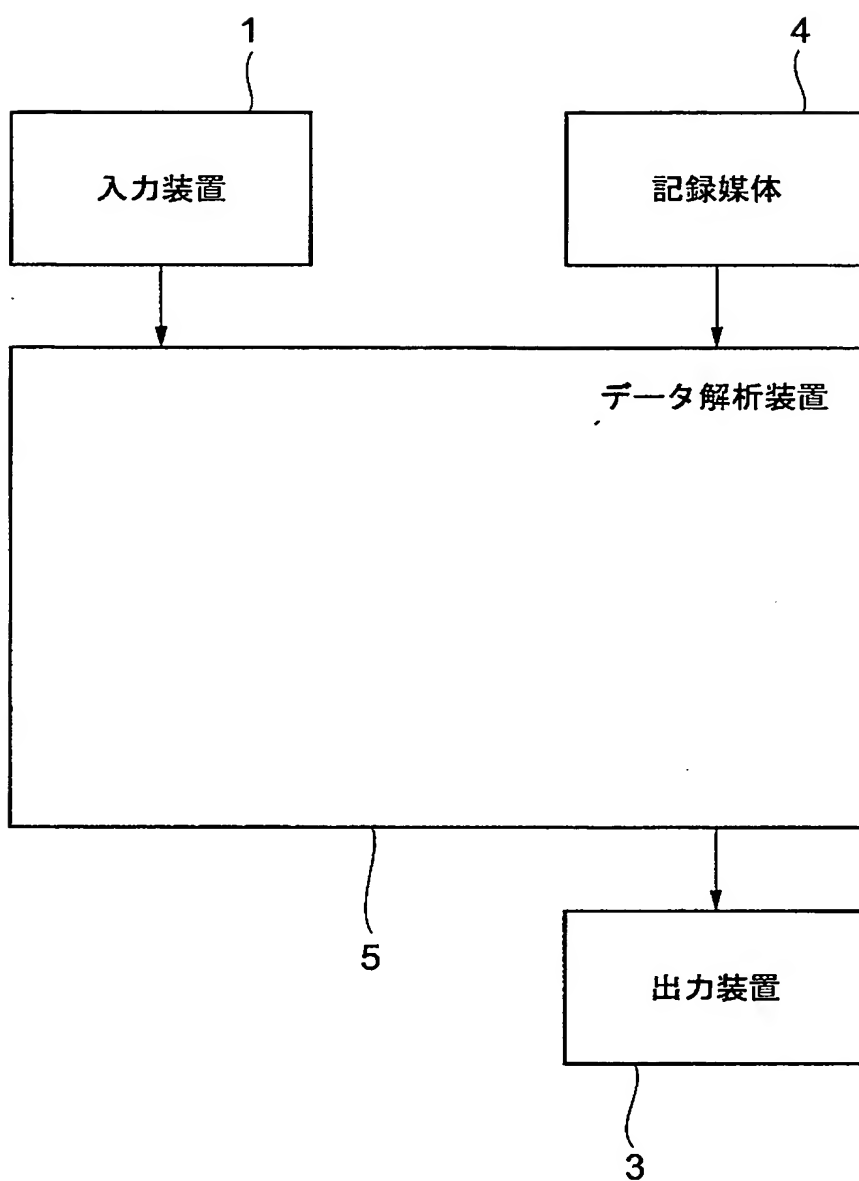
【図 2】



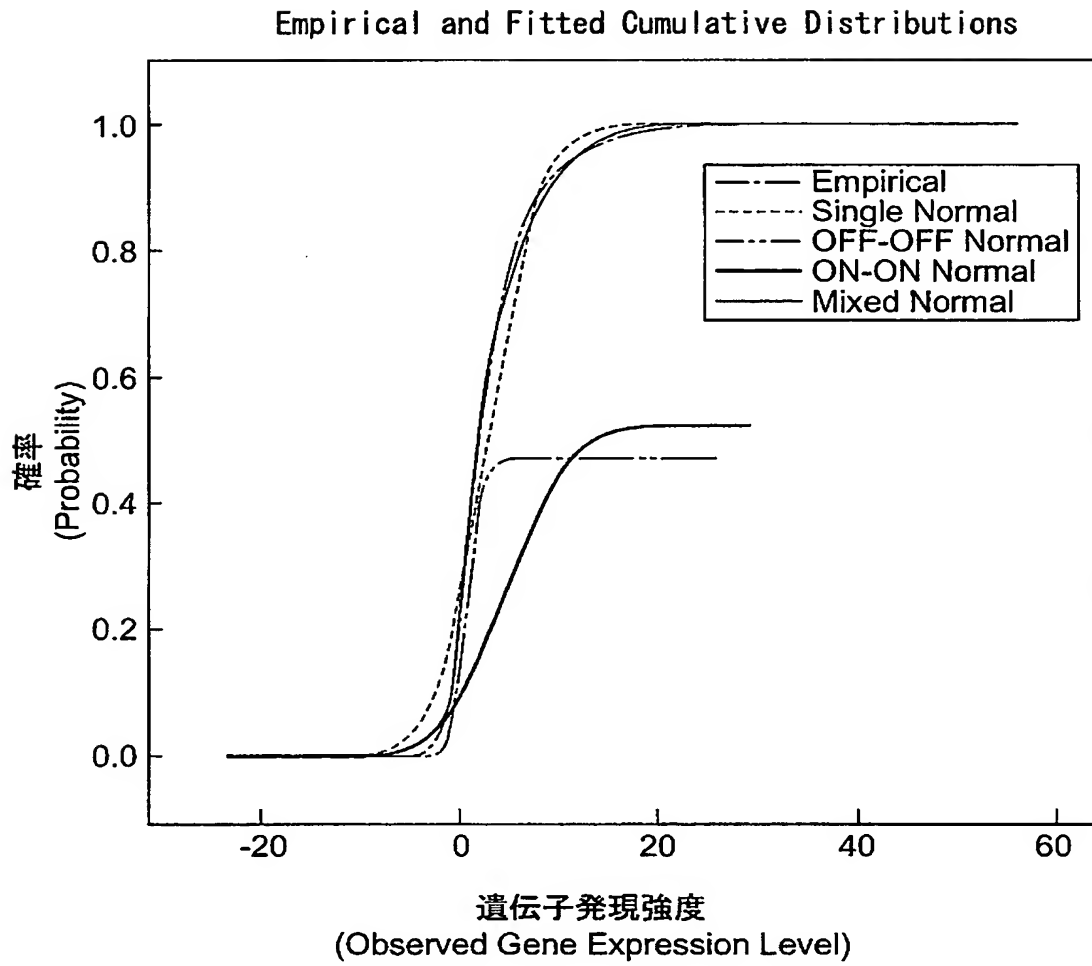
【図 3】



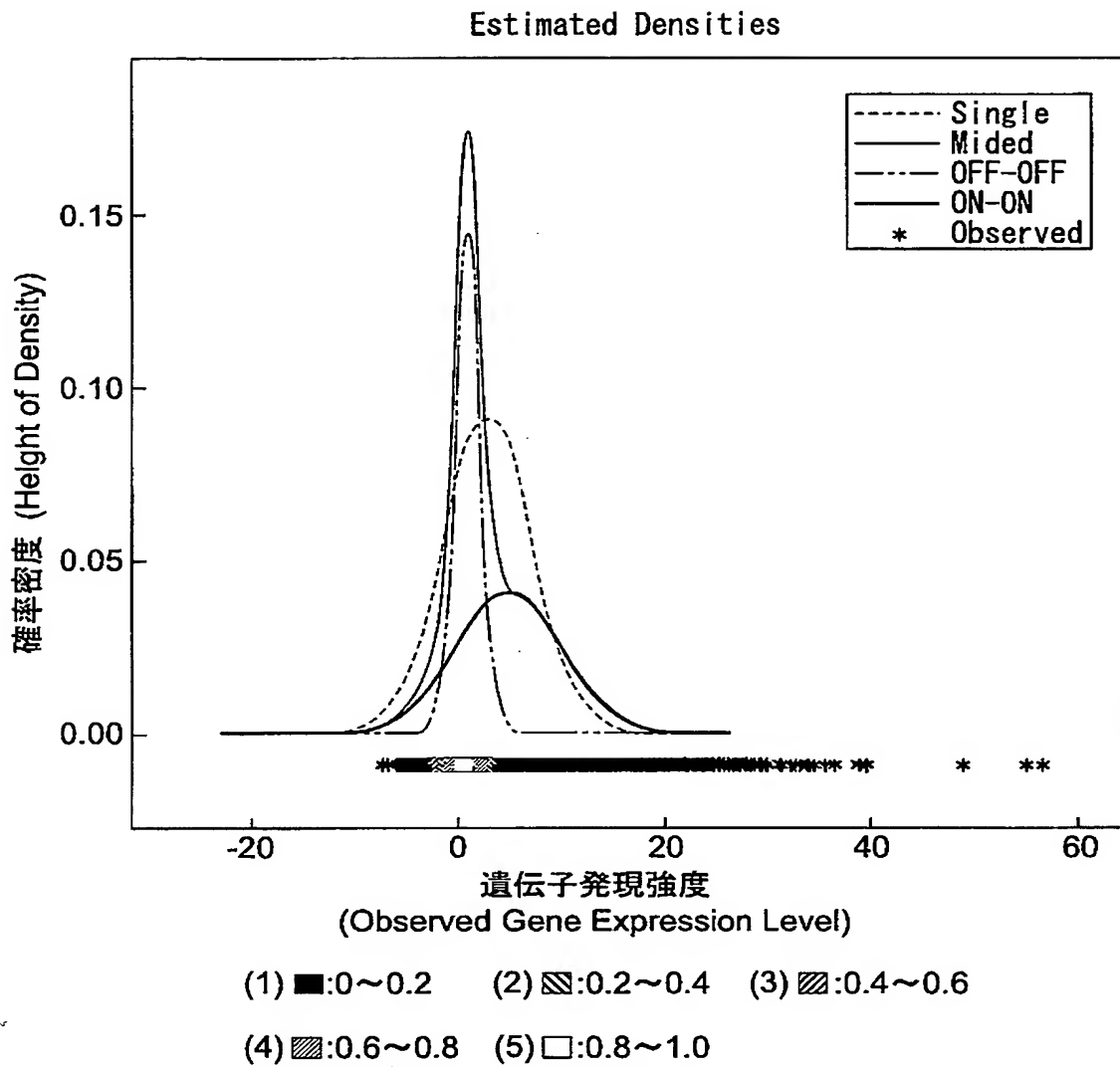
【図 4】



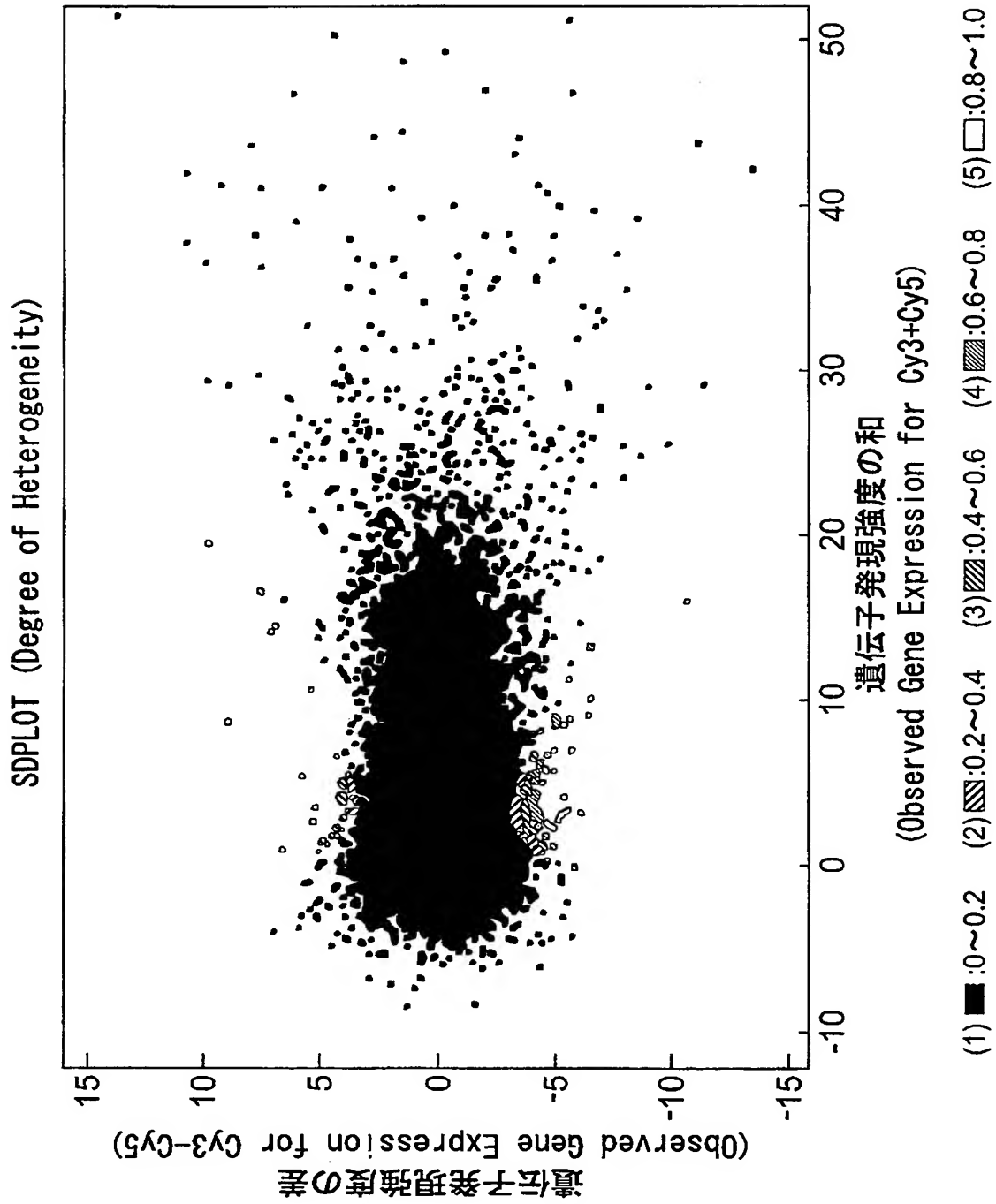
【図 5】



【図 6】



【図 7】



【書類名】 要約書**【要約】**

【課題】 マイクロアレイデータを用いた解析の精度および効率を高めるために、遺伝子の発現状況に関する真の信号と測定誤差の分離を行い、チャンネル毎に遺伝子が発現している確率を推定する。

【解決手段】 本発明の遺伝子発現状況推定システムは、マイクロアレイデータを入力する入力装置 1 と、プログラム制御により動作するデータ解析装置 2 と、出力装置 3 とを含む。データ解析装置 2 は、入力装置 1 から与えられた遺伝子発現強度データを用いて、混合正規分布のコンポーネント毎の分布パラメータおよび混合比パラメータを推定するパラメータ推定手段 2 1、2 2 と、推定された各パラメータを用いて、チャンネル毎に遺伝子の発現状況に関する事後確率を計算する事後確率計算手段 2 3 を有し、計算された事後確率は出力装置 3 に出力する。

【選択図】 図 2

特願 2 0 0 3 - 2 7 5 9 8 3

出 願 人 履 歷 情 報

識別番号

[0 0 0 0 0 4 2 3 7]

1. 変更年月日

1 9 9 0 年 8 月 2 9 日

[変更理由]

新規登録

住 所

東京都港区芝五丁目 7 番 1 号

氏 名

日本電気株式会社

特願 2 0 0 3 - 2 7 5 9 8 3

出 願 人 履 歴 情 報

識別番号

[5 0 3 0 7 7 1 6 5]

1. 変更年月日

2 0 0 3 年 2 月 2 6 日

[変更理由]

新規登録

住 所

広島県廿日市市宮園9丁目1の7

氏 名

大瀧 慈

特願 2 0 0 3 - 2 7 5 9 8 3

出 願 人 履 歴 情 報

識別番号

[5 0 0 5 3 5 3 0 1]

1. 変更年月日

2 0 0 0 年 1 1 月 2 0 日

[変更理由]

新規登録

住 所

東京都中央区八丁堀二丁目 2 6 番 9 号 グランデビルディング

氏 名

社団法人バイオ産業情報化コンソーシアム